Revealing Hidden Community Structures and Identifying Bridges in Complex Networks: An Application to Analyzing Contents of Web Pages for Browsing

Faraz Zaidi, Arnaud Sallaberry, Guy Melançon

Revealing Hidden Community Structures and Identifying Bridges in Complex Networks: An Application to Analyzing Contents of Web Pages for Browsing

Faraz Zaidi
CNRS UMR 5800 LaBRI &
INRIA Bordeaux - Sud Ouest
351, cours de la Libération
33405 Talence cedex, FRANCE
faraz.zaidi@labri.fr

Arnaud Sallaberry
CNRS UMR 5800 LaBRI &
INRIA Bordeaux - Sud Ouest & Pikko
351, cours de la Libération
33405 Talence cedex, FRANCE
arnaud.sallaberry@labri.fr

Guy Melançon
CNRS UMR 5800 LaBRI &
INRIA Bordeaux - Sud Ouest
351, cours de la Libération
33405 Talence cedex, FRANCE
guy.melancon@labri.fr

Abstract

The emergence of scale free and small world properties in real world complex networks has stimulated lots of activity in the field of network analysis. An example of such a network comes from the field of Content Analysis (CA) and Text Mining where the goal is to analyze the contents of a set of web pages. The Network can be represented by the words appearing in the web pages as nodes and the edges representing a relation between two words if they appear in a document together. In this paper we present a CA system that helps users analyze these networks representing the textual contents of a set of web pages visually. Major contributions include a methodology to cluster complex networks based on duplication of nodes and identification of bridges i.e. words that might be of user interest but have a low frequency in the document corpus. We have tested this system with a number of data sets and users have found it very useful for the exploration of data. One of the case studies is presented in detail which is based on browsing a collection of web pages on Wikipedia [1].

1. Introduction

Information Analysis (IA) or more precisely Content Analysis (CA) is an active area of research where the goal is to explore and analyze contents of a set of documents in order to discover patterns and hidden knowledge [25]. Weare provides a good overview of the challenges presented to the CA research community by the World Wide Web [30]. Document Content Visualization Systems can be used as a tool for CA where the goal is to represent textual contents of a set of documents in a visual form so as to facilitate the process of mining and discovering patterns in a collection of documents [21] [15].

A typical application of CA in the domain of web is the analysis of the set of web pages browsed by a user in order to find required information. While browsing a web page having external links; it is imperative for the users to browse each and every external link if further information is required. This task is not only time consuming but makes it difficult for the users to relate contents of web pages to each other. Moreover apart from going over a single web page, most of the time, users tend to collect a set of web pages rather than a single web page to obtain information [31]. In case of browsing, it means that a user would explore the links in the selected pages further to extract more information.

As an example consider browsing the web page 'CAC 40' on the Wikipedia encyclopedia. CAC 40 is a benchmark french stock market index which represents a capitalization-weighted measure of the 40 most significant values among the 100 highest market caps on Euronext Paris [2]. There are many external links on this page and users searching for more details would use these external links to look for further information. Usually they will go through the web pages one at a time to find more information which makes it difficult to relate what they have already found as information.

Ideally the users want a system that collects pages referred in the initial web page and display their contents in a manner that groups the pages together based on the content. This would give the user an idea about the sub-topics that revolve around the major topic. For example, a quick look on the CAC 40 Wikipedia page would suggest the other related indices like 'CAC Next 20', 'CAC Mid 100' and 'CAC Small 90' as possible topics related to 'CAC 40'. All these indices represent the highest equities on Euronext Paris where CAC Small being the lowest of these all.

Moreover the users would like to see how these sub-topics are related to each other. Words appearing in more than one document can play a role of bridges between these groups thus creating a link between words from two different groups. Nodes organized into circles in Fig. 1 cor-

1. http://en.wikipedia.org/wiki/Main_Page

2. http://en.wikipedia.org/wiki/CAC_40

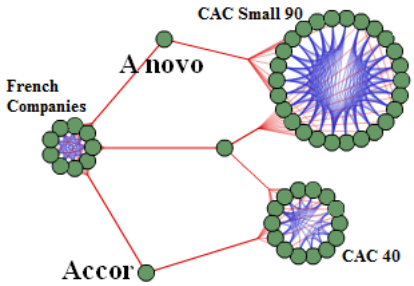

Figure 1. Web Pages and Bridges

respond to keywords extracted from a same set of document. Clearly, the figure shows that documents roughly organize into three different subgroups of topics, labeled as "French companies", "CAC 40" and "CAC Small 90". In other words, Fig. 1 shows the "community structure" of web pages [14] [2]. Additionally, a few nodes have been isolated and bridge these subgroups, further indicating topics that link the different subgroups. The bridging nodes thus sit out of the community structure (the structural hole as Burt names it [9]) and act as broker between communities.

Our starting point is thus the network of links between words extracted from web pages. As we shall see in the coming sections(3), the plain network we obtain after the extraction process is actually quite complex and its structure forbids to use it as a graphical representation as such. The problem we address in this paper is to reveal hidden community structures in complex networks. Our approach also allows us to identify the bridges within these networks. The uncovered structure can then be used to build a graphical representation of the whole network to help visualize and interactively explore these relationships.

We represent the contents of the web pages through a graph or a network, where the nodes represent the words appearing in web pages and the edges showing that two words appear together at least once on a web page. Clustering and Visualization of a network constructed this way is a challenging problem because having a closer look at the network properties, we find that these networks have the properties of small world [20] and scale free [3] structures.

A small world network can be defined by two structural parameters: the average path length [20] and the clustering coefficient [29] of nodes. The path length refers to the minimum number of edges traversed to go from a node A to node B. The average path length is the average calculated for all pair of nodes in a network. The clustering coefficient of a node is defined as the ratio between the total number of connections among the neighbors of that node to the total number of possible connections between the neighbors. A high clustering coefficient means that the neighborhood shares lots of connections among themselves and thus form a community.

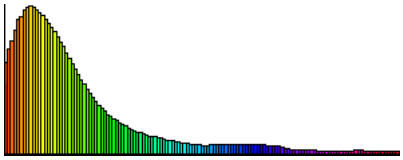

Figure 2. Plot of the node degree distribution : degrees appears on the x-axis and the frequencies associated on the y-axis.

A scale free network is a network in which the connectivity of nodes follow a scale-free power-law distribution [3]. This means that there are a few nodes that have a very high number of connections(degree) and lots of nodes are connected to a few nodes. The degree distribution of the network taken as an example is shown in figure 2 where the long tail like structure represents the nodes with high number of connections.

The most common networks having these properties of both scale free networks and small world networks are the Internet Movie Database and the Coauthoring Network [16]. For the rest of the paper, we use the term complex network to refer to a network which has both small world and scale free properties. A Cooccurrence network is another example of a complex network. This network is built when one considers a book or queries to a search engine, one can construct a co-occurrence graph of words if they appear in the same sentence or on the same page [10]. This is the graph that we construct considering the web pages browsed by a user while exploring for information (see section 3 for details).

We know that due to the scale free property, it is difficult to cluster a network or visualize communities within the network structure [1]. This is due to the fact that a few nodes have a very high degree and they create links between the underlying community structure hiding the communities. Figure 3 shows the co-occurrence network obtained by browsing pages on the Wikipedia encyclopedia starting from the CAC 40 page (see section 3 for details). We have used a graph drawing algorithm proposed by Frick [13] implemented in Tulip software [3]. This is force directed algorithm which is well suited to our problem as it puts the nodes that are densely connected to each other closer hence making it easier to locate the community structures. But from the figure it is quite evident that in the presence of very high degree nodes it is very hard to identify different communities visually.

The proposed system addresses two main problems in the analysis of complex networks. First is revealing the community structures hidden in the network through simplification of the graph and clustering. As an example we consider the co-occurrence graph of words extracted from a set of web pages. In the discussed example of CAC 40 is the major

3. http://www.tulip-software.org/

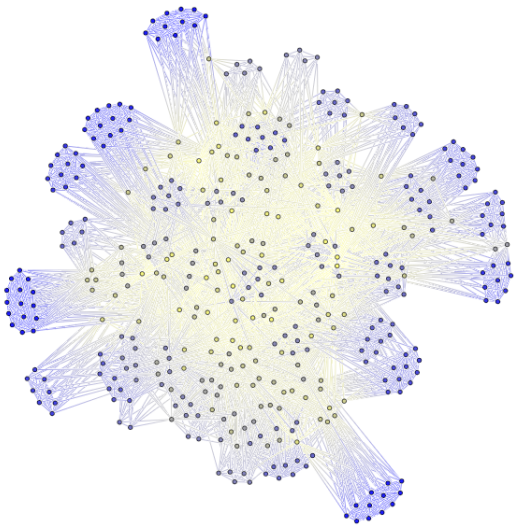

Figure 3. Co-occurrence Network of Words: Pages Browsed from CAC 40 Wikipedia web page

topic but the user might be interested in other indices such as CAC Small 90 etc.

The second is identification of words that are interesting from a user perspective to study the relationship between clusters but have a low frequency in the entire document corpus. These words can help us to uncover hidden information and discover relationships that are not apparent in the original network or difficult to locate. For example a company linking web pages from CAC Next 20 and CAC 40 might be a good candidate of a company that has interests in equities from CAC Next 20 and CAC 40. This company name should be represented separately as a bridge that the user can find easily without having to go through the two web pages and deduce this relationship.

The major contributions of the proposed system are: a methodology to cluster complex networks based on duplication of nodes, a method to focus on Bridges that are interesting from a user perspective, discover hidden (or not so obvious) relationships between subtopics based on these Bridges. We use Micro/Macro graph layout algorithms [4] to visualize the final network which helps us to develop an overall picture of the distribution of the contents of the collected set of web pages.

The rest of the paper is organized as follows: In Section 2 we present the related work. We describe the data set used as an example in section 3 and the proposed system in section 4. Section 5 discusses the results that we obtained by the application of our framework on the sample data set. Section 6 contains the conclusion and future research prospects, advancements and ameliorations possible to the current system.

2. Related Work

Document Content Visualization has been studied in details by various researchers and different visualization systems have been proposed such as [15] [11]. Most of these systems are useful to identify the key words in a document collection. They use the classical techniques to calculate the relationships between documents like the tf-idf score [24] which makes it difficult to focus on low frequency words that appear in only a few documents.

Web search results visualization which is different from web browsing visualization is also an active area of research where most of the research has been directed towards two objectives. One is, towards showing the connection between the user query and the resulting web pages [22] and two, effective organization and clustering of search results [6]. None of these systems perform content analysis as their primary task is to help users find web pages relevant to their query. Some of these systems perform clustering based on the content of the web pages [33] but none of these focus on showing the bridges that create links between these documents.

Clustering and Visualization of a network or graph having small world and scale free properties at the same time, to the best of our knowledge, has not attracted much attention in the clustering domain. Boutin et al. [7] tried to address this issue by introducing a filtering method based on user-focus. This technique extracts a tree-like graph so that the resulting structure can be drawn using any force directed algorithm leaving the final drawing easily readable. One of the drawbacks of this system is that the user has to choose an initial entry point to filter the graph. The system we propose require no such information. Moreover since edges are removed to simplify the structure of the network, important information can be lost. We preserve the original network without removing any edges or nodes thus no information loss occurs and the user is free to navigate in the entire network all the time.

Girvan et al. [14] proposed a clustering algorithm for complex networks based on the betweenness centrality. This algorithms performs well but produces a hierarchical clustering which is not well suited to browsing problem addressed in this paper. Moreover it does not cater the problem of reducing the inter-cluster edges thus making it difficult for visualization.

Methods have been proposed to cluster and visualize scale Free networks based on filtering [23] [19] of nodes or edges but some loss of information occurs at the same time. For example [1] [19] propose a recursive pruning method called K-core which is a method to simplify large scale free networks for visualization. It progressively allows the detection of central nodes in the network but the grouping is solely based on the topology and does not reflect the similarity of the nodes. [23] proposed a method based on

the on Minimum Spanning tree(MST) , where the goal is to construct a MST of the network thus reducing the network from a graph to a tree. This essentially requires deletion of edges and loss of information does occur.

Efficient algorithms to cluster and visualize small world networks have been proposed like [2] [27]. These systems perform well if the topology of the network follows small world properties but fail to perform in the presence of scale free properties. This is due to the fact that in a scale free network, a few nodes dominate the entire networks connections and makes it difficult to identify the communities.

3. Collection and Preprocessing of Data

The data set that we have used for experimentation is the collection of key words found in the Meta tag of a set of web pages. These web pages were collected starting from the page CAC 40 on Wikipedia [4]. All the pages in the 'See Also' section were further explored and the process was repeated for links up to depth 3. A total of 50 web pages were collected this way. The choice of selecting 50 pages was influenced by the study [17] which shows that users will try a new search after browsing at most 30 web pages in case of searching a web page on the Internet. The total number of words collected after the removal of stop words (like 'for', 'the' etc.) was 412.

The entire data set can be represented by 3 tuples. One for the Words (words), second for the documents (document title, hyper link) and the third representing relationship between the documents and the words, Relationship(Document title, words).

From this data set, two different graphs can be constructed. A graph of Web page-Word and a Word-Word graph. In a Web page-Word graph, the nodes represent the web pages and the words where an edge between a web page and a word represents that the word appears in that web page. This graph by construction forms a bi-partite graph where there are no edges between words and similarly there are no edges between the web pages. We use this graph to find the words that appear in many web pages as the degree of a word represents the number of web pages it appears in.

The other graph is the word-word graph which we eventually use for visual analysis of the network. The nodes represent the words and an edge between two words represent that they appear together in at least one web page. This final graph contains 412 nodes and 5817 edges. The graph is shown in Fig. 3. An important observation about this graph is that the words that appear in a single web page would be connected to each other thus forming a clique. Looking carefully at Fig. 3, the set of nodes that form a group and are densely connected to each other most probably belong to the same web page.

4. http://en.wikipedia.org/wiki/List_of_french_companies

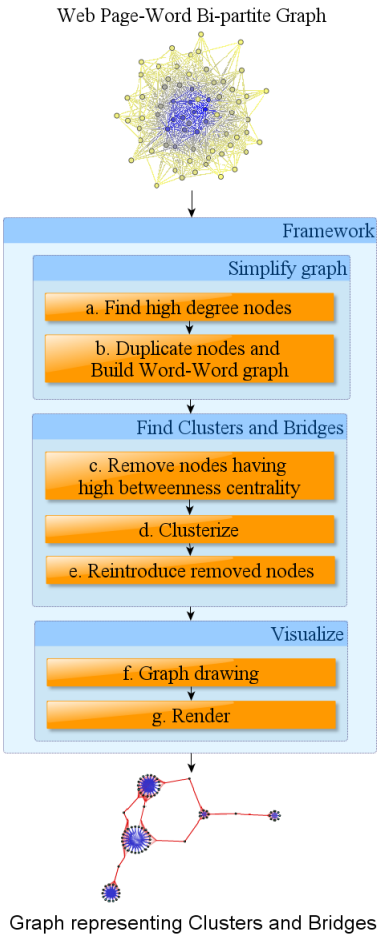

Figure 4. Proposed Framework

4. Framework of Proposed System

The inspiration to the proposed system comes from the fact that if we are somehow able to reduce the complex graph (a graph with both scale free and small world properties) to a small world graph, clustering and visualization of a small world graph is much easier especially if the resulting graph has communities with high intra community edges and low inter community edges [2].

To reduce the complex graph to a small world graph without scale free property we duplicate nodes that have a very high degree thus leaving us with only a small world network. In our case, duplicating the high degree nodes means that the words that are present in a majority of documents are duplicated such that they are assigned a new identity in each and every document they appear. Thus they are treated as words that appear only in a single document. This is done so in order to reduce the inter-cluster edges which in turn, results in a more readable visualization of the network. Thus revealing the community structure visually.

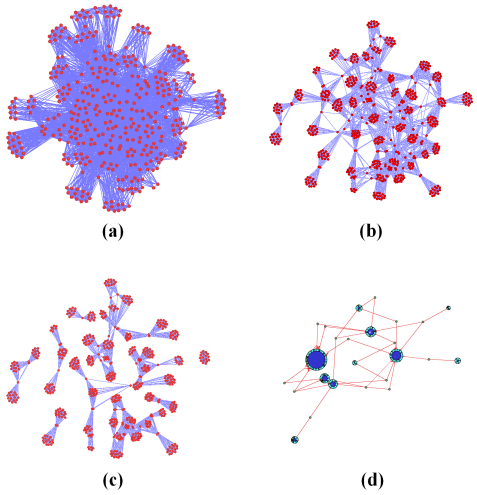

Figure 5. (a) Word-Word Graph constructed from browsing CAC 40 and related web pages (b) Graph after node duplication (c) Graph after removing bridges (d) Graph with Clusters and Bridges using proposed visualization

Then we use the Betweenness Centrality introduced in [12] (see also [8] for implementation) to identify the nodes that lie between communities of words representing the small world structure. These are the words that are present in a few documents only and play the role of bridges between web pages, i.e. they link different web pages. After identifying these words, we remove them temporarily further simplifying the entire network to reveal disconnected or loosely connected communities. A clustering algorithm can be applied to clusterize this network.

Once the clusters are found, the words that were initially duplicated might found themselves in the same cluster. We remove the duplicated nodes within clusters so as to keep a single copy of the duplicated nodes. Then we reintroduce the nodes having high betweenness centrality that were removed temporarily and we identify them as Bridges.

Finally the network of clusters and Bridges is drawn using a graph drawing algorithm. We associate a different color to identify the nodes that are duplicated in the network so as to show users the nodes that are present in other clusters as well. Thus we have nodes that have two different colors (see Fig. 7(a)), representing nodes that appear only once or are duplicated. A simple interaction by clicking a duplicated node is introduced to trace the presence of a duplicated node in the entire network by associating a third color (see Fig. 7(b)). The following sections discuss our framework in detail.

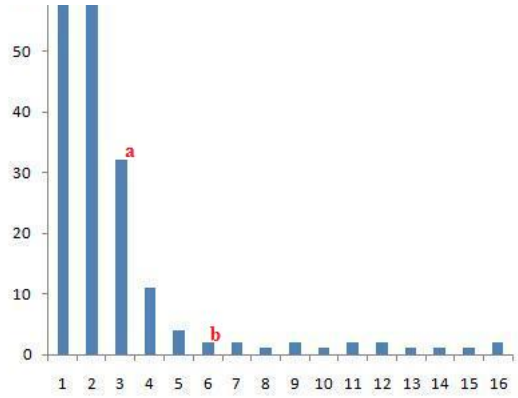

Figure 6. Number of Web pages on the x-axis and Number of Words on the y-axis

4.1. Using Scale Free structure to find cut off point and duplicate nodes

In order to find the words to be duplicated, we use the bi-partite graph of words and documents as described in section 3. Once we have the Web page-Word graph, we need to identify the words that are present in many documents. The degree of the nodes in this graph represents the number of documents a word appears in. Fig. 6 shows the frequency distribution of the words and the web pages. The x-axis represents the number of documents and the y-axis represents the frequency of words. For example the point 'a' in the Fig. 6 means that there are just a little over 30 words that appear in exactly three web pages.

Since the idea is to duplicate words that appear in many documents, we need to find the proper definition of what many document means for this network. Looking at Fig. 6 we calculate the slope of every two consecutive points. At point b the slope becomes equal to zero. This gives us a heuristic which suggests that as the slope becomes zero or close to zero (values of -1 or -2) this point can be considered as the cutoff point. In the given example, it turns out to be 6, meaning that all the words that appear in 6 or more documents must be duplicated. Although this heuristic provides a good starting point for the system, the user is free to manually choose a value for the degree a part from which the nodes would be duplicated. Lower the value chosen, higher would be the number of words being duplicated and the eventual word-word graph would become more disconnected.

An example of a word that might be duplicated is the word 'France' since CAC 40 is an index for french stock exchange, it is quite obvious to find this word in many documents. This step introduces new nodes in the word-word graph but keeps the number of edges exactly the same. Thus the graph gets simplified as shown in Fig. 5(a) and 5(b).

4.2. Iterative removal of Nodes with high Betweenness Centrality

Once we have the word-word graph with duplicate nodes, we calculate the betweenness centrality of the nodes which is a metric proposed by Freeman [12]. It calculates the relative importance of nodes within a network by calculating the shortest paths between pairs of nodes. Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not. This metric is a good representation of the nodes that play the role of connecting different communities and thus helps us in identifying the bridges.

Girvan et al. [14] have used a clustering algorithm which is based on a modified form of this metric. They calculate edge betweenness based on the same principal where they find edges that are central to a network. Then they iteratively remove the most central edge in the network and recalculate the edge betweenness until no more edges are left.

Since our goal is to find the bridges we apply the same method on nodes to identify the bridges. We calculate the betweenness centrality of the nodes, remove the node with the highest betweenness centrality and repeat this process a certain number of times. Girvan repeated the process until there were no edges left as the goal was to produce a hierarchical clustering. We use a heuristic to determine the number of iterations which is based on the total number of documents used for extraction of words and the number of bridges we want to see between groups of documents. For the given example we choose 15 as it would give us a bridge for nearly every three documents.

$$Number\ of\ Iterations = \lfloor Number\ of\ Documents/3 \rfloor \quad (1)$$

Where $\lfloor\rfloor$ represents the ceiling function.

The user is free to choose any value depending on the requirements, higher the number of iterations and higher would be the number of bridges and smaller would be the size of clusters. Fig. 5(c) represents the graph after the removal of nodes with high betweenness centrality.

4.3. Finding Communities through Clustering

A clustering algorithm is required to find clusters in this processed graph. Most of the clustering algorithms require some parameters as input such as number of clusters to be found, initial centroids, a threshold etc [5] [18] [32]. Since we want to avoid bothering the user with the requirement of this additional information, we prefer an algorithm that does not require any information from the user. Strength clustering proposed by Auber et al. [2] is a good choice as it requires no such parameter. It is well suited to our problem as it was particularly designed to find communities in small world graphs.

Once the clusters are found, each and every cluster is scanned for nodes that were duplicated and found themselves in the same cluster. We remove this node duplication within a cluster and keep a single instance of a duplicated node within a cluster.

4.4. Reintroducing Nodes with High Betweenness Centrality and Identification of Bridges

The next step after clustering of the nodes is to re-introduce the nodes that were earlier removed due to high betweenness centrality. These nodes are considered to be the Bridges as they are responsible for connecting different clusters. Keeping in view that the words that were present in many web pages were duplicated and thus their degree was reduced, the nodes having high centrality in this final network are words that appear in a few web pages only. These words are important from the user perspective as they link web pages and might play an important role to understand the relationship between webpages.

4.5. Visualization of Clusters and Bridges

The final graph is visualized using a variation of the Micro/Macro graph layout algorithms [4]. Providing a detailed account of the layout algorithm we designed is out of the scope of this paper. Roughly speaking, once communities have been identified they are treated as metanodes and fed into a force-directed layout, as to benefit from their aesthetics and readability [26] [28]. The GEM algorithm [13] has been adapted as to take node size into account in order to reflect the size of a cluster (number of words it contains).

To draw the nodes representing words within a cluster we use a simple placement strategy. They are placed uniformly on the circumference of the circle surrounding the cluster. The ordering of nodes must however be decided in order to minimize the number of edge crossings it might introduce in the final drawing. Two different colors are associated to the nodes within a cluster. (Light green and dark green as shown in Fig. 7.) The light green color represents the nodes that are duplicated throughout the network. An important interaction is clicking on a duplicated node, which highlights all of its instances in the entire network as shown in Fig. 7(b) using a different color (Pink) and size. This is to help locate the duplicated instances of a node in the network.

5. Results and Discussion

Figure 7 shows a small part isolated from Fig. 5(d). Fig. 7(a) shows the three clusters represented by circular structures having many small nodes and the bridges which are labeled 'A novo', 'AXA', 'Accor' and 'number'. The first three represent french companies and the bridge 'number'

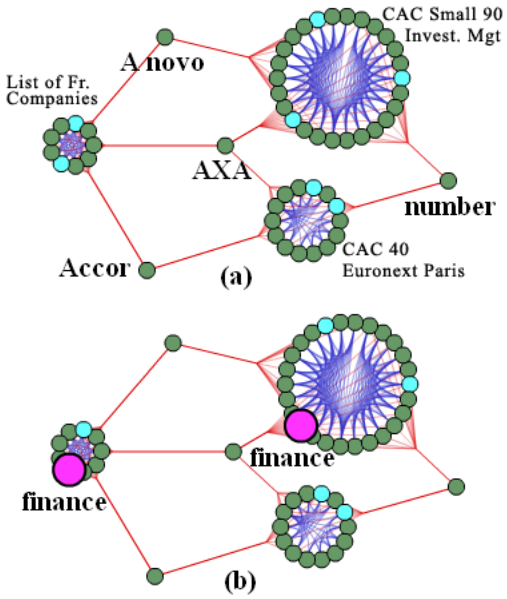

Figure 7. (a) A small part isolated from Fig. 5(d) showing Titles of Webpages clustered together and Bridges(b) Duplicated nodes highlighted after selection

is a noise. the clusters are associated with titles of the web pages, which are CAC Small 90, Investment Management in a cluster, CAC 40 and Euronext Paris in another cluster and the third cluster containing only the document List of French Companies. In Fig. 7(b), the word Finance was selected in a cluster, which is a duplicated node and is present in two clusters.

CAC 40 and Euronext Paris represents the stock exchange of Paris, where CAC 40 is an index based on the 40 biggest equities of France. The web page CAC Small 90 is an index representing the 90 biggest equities after CAC 40, CAC Next 20 and the CAC Mid 100. The page Investment Management is about the companies that are interested in investment. The page List of french Companies contains a list of all the french companies.

Looking at the clustering, The first cluster, which contains the CAC 40 and the Euronext Paris, it is obvious that these two pages find themselves in the same cluster as they both represent the Paris Stock Exchange. 'AXA' and 'A Novo' are two companies where AXA is listed in the CAC 40 and A Novo is listed in CAC Small 90. Finding CAC Small 90 with Investment Management makes sense as AXA is a french company interested in investments and targets companies in the CAC Small 90 as a possible investment opportunity where A novo is an example of a possible future investment. Similarly the relation between List of french companies and CAC 40 through Accor suggests that Accor is a french company listed in CAC 40.

All this analysis is a direct result of the visual representation of the network. Clusters group things that have similarity based on the content and Bridges are responsible for creating relationships between these clusters giving an overall understanding of the data set.

6. Conclusions and Future Work

In this paper we have presented a system to visualize and explore complex networks revealing clusters and detecting bridges in a set of web pages. The system was tested with several examples considering the web pages browsed by users. The in-house informal tests with different users indicate that the system was found to be very useful to develop and overall understanding of the collection of web pages. The identification of the subtopics revolving around the primary search topic was a direct result of the clustering. The identification of the words that play the role of bridges between these different subtopics was also found to be very useful.

The system was tested with small data sets as the web browsing on a single topic does not require to evaluate hundreds of web pages at the same time. Similarly the size of documents was not very huge as web pages usually have a very limited size as compared to books, newspapers etc. As part of the future work, we would like to test the system to visualize web search results as compared to browsing. We would also like to ameliorate the system to incorporate the exploration of complex networks of large sizes such as the Internet Movie Database. We also plan to introduce more interactions to facilitate the user navigation like deleting nodes, dragging nodes from one cluster to the other etc.

Acknowledgment

This project was partly funded through the ANR project RNTL FIVE 06 TLOG 12. The project aims to develop visual interfaces to exploit effectively the information collected as a result of text mining of web pages and introduce modes of interaction to help analysts monitor and extract information. We would like to thank the members of the organization Pikko [5], Mr. Emmanuel Pellegrin who helped us analyze and interpret the data and Mr. Guillaume Aveline who helped us to collect the data.

References

[1] Jose Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the visualization of large scale networks. *Advances in Neural Information Processing Systems*, 18:41–50., 2006.

[2] D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. Multiscale visualization of small world networks. In *INFOVIS '03: Proceedings of the IEEE Symposium on Information Visualization*, pages 75–81, 2003.

5. http://www.pikko-software.com/

[3] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[4] Michael Baur and Ulrik Brandes. Multi-circular layout of micro/macro graphs. In Seok-Hee Hong, Takao Nishizeki, and Wu Quan, editors, *Graph Drawing*, volume 4875 of *Lecture Notes in Computer Science*, pages 255–267. Springer, 2007.

[5] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002. http://www.accrue.com/products/rp_cluster_review.pdf.

[6] Nicolas Bonnel, Vincent Lemaire, Alexandre Cotarmanac'H, and Annie Morin. Effective organization and visualization of web search results. In *Proceedings of the 24th IASTED International Multi-Conference Internet and Multimedia Systems and Applications*, February 2006.

[7] Francois Boutin, Jérôme Thievre, and Mountaz Hascoët. Focus-based filtering + clustering technique for power-law networks with small world phenomenon. In *VDA'06: Visual Data Analysis - SPIE-IS&T Electronic Imaging*, volume 6060, pages 001–012. SPIE P., U.S., 2006.

[8] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[9] Ronald S. Burt. *Brokerage and Closure*. Oxford University Press, 2005.

[10] R. Ferrer I Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London*, B268(1482):2261–2265, November 2001.

[11] B. Fortuna, D. Mladenic, and M. Grobelnik. Visualization of text document corpus. *Informatica Journal*, 29(4):497–502, 2005.

[12] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[13] A. Frick, A. Ludwig, and H. Mehldau. A fast adaptive layout algorithm for undirected graphs. In *GD '94: Proceedings of the DIMACS International Workshop on Graph Drawing*, pages 388–403, London, UK, 1995. Springer-Verlag.

[14] Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:8271–8276, 2002.

[15] M. Grobelnik and D. Mladenic. Visualization of news articles. *Informatica Journal*, 28, 2004.

[16] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. In *Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN), LNCS*, volume 1, 2004.

[17] iProspect. iprospect's search engine user attitudes. survey results. white paper, 2004.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[19] Yuntao Jia, Jared Hoberock, Michael Garland, and John Hart. On the visualization of social and other scale-free networks. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1285–1292, 2008.

[20] Stanley Milgram. The small world problem. *Psychology Today*, 1:61–67, May 1967.

[21] Kimberly A. Neuendorf. *The Content Analysis Guidebook*. Sage Publications, Inc, December 2001.

[22] T.N. Nguyen and J. Zhang. A novel visualization model for web search results. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):981–988, Sept.-Oct. 2006.

[23] Niina Päivinen. Clustering with a minimum spanning tree of scale-free-like structure. *Pattern Recogn. Lett.*, 26(7):921–930, 2005.

[24] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[25] VinhTuan Thai, Siegfried Handschuh, and Stefan Decker. Ivea: An information visualization tool for personalized exploratory document collection analysis. In Manfred Hauswirth, Manolis Koubarakis, and Sean Bechhofer, editors, *Proceedings of the 5th European Semantic Web Conference*, LNCS, Berlin, Heidelberg, June 2008. Springer Verlag.

[26] Ioannis G. Tollis, Giuseppe Di Battista, Peter Eades, and Roberto Tamassia. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

[27] F. van Ham and J.J. van Wijk. Interactive visualization of small world graphs. In *INFOVIS 2004. IEEE Symposium on Information Visualization*, pages 199–206, 2004.

[28] D. Wagner and M. Kaufmann. *Drawing Graphs, Methods and Models*, volume 2025 of *LNCS*. Springer, 2001.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

[30] Christopher Weare and Wan-Ying Lin. Content analysis of the world wide web: Opportunities and challenges. *Social Science Computer Review*, 18(3):272–292, August 2000.

[31] Yi-fang Brook Wu, Latha Shankar, and Xin Chen. Finding more useful information faster from web search results. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 568–571, New York, NY, USA, 2003. ACM.

[32] Illhoi Yoo and Xiaohua Hu. A comprehensive comparison study of document clustering for a biomedical digital library medline. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 220–229, New York, NY, USA, 2006. ACM Press.

[33] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2004. ACM.